ON THE SPARSE BAYESIAN LEARNING OF LINEAR MODELS

CHIA CHYE YEE AND YVES F. ATCHADÉ

(Dec. 2014)

ABSTRACT. This work is a re-examination of the sparse Bayesian learning (SBL) of linear regression models of Tipping (2001) in a high-dimensional setting. We propose a hard-thresholded version of the SBL estimator that achieves, for orthogonal design matrices, the non-asymptotic estimation error rate of $\sigma\sqrt{s\log p}/\sqrt{n}$, where n is the sample size, p the number of regressors, σ is the regression model standard deviation, and s the number of non-zero regression coefficients. We also establish that with high-probability the estimator identifies the non-zero regression coefficients. In our simulations we found that sparse Bayesian learning regression performs better than lasso (Tibshirani (1996)) when the signal to be recovered is strong.

1. INTRODUCTION

High-dimensional variable selection has become an important topic in modern statistics. The least absolute shrinkage and selection operator (lasso) of Tibshirani (1996) is probably the most widely used method for this problem and has span an extensive literature (see e.g. the monograph Bühlmann and van de Geer (2011)). Despite its success, the method has many shortcomings. For instance choosing the right amount of regularization remains a difficult and computer-intensive issue for many models. In parallel to the frequentist approach, Bayesian variable selection for high-dimensional problems has also generated a large literature (see for instance O'Hara and Sillanpää (2009) and the reference therein). But most Bayesian variable selection methods often

C. C. Yee: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* chye@umich.edu.

Y. F. Atchadé: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* yvesa@umich.edu.

lead to intractable posterior distributions that require a heavy use of Markov Chain Monte Carlo simulation. Between these two well-established frameworks lies an empirical Bayes alternative known as sparse Bayesian learning (SBL, Tipping (2001)), which has received much less attention in the statistical literature.

This paper is a re-examination of the SBL for linear regression in a high-dimensional setting. An interesting question is whether the SBL procedure recovers the sparsity structure of underlying signals. This problem was considered by Wipf and Rao (2004) which establishes that in the noiseless setting the SBL indeed recovers the sparsity structure of the regression coefficients. However the method behaves differently in a noisy setting. For orthogonal design matrices, we show that the SBL indeed produces a sparse solution of the regression coefficients, but does not in general recover the sparsity structure of the regression coefficients. To remedy this limitation we propose a hard-thresholded version of the SBL estimator. We show that with high probability the thresholded estimator achieves the same estimation error of $O(\sigma\sqrt{s\log(p)/n})$ as lasso, where n is the sample size, σ is the regression model standard deviation, p the number of regressors and s the number of non-zero regression coefficients. Furthermore we show that with high probability this thresholded estimator recovers the sparsity structure of the regression coefficients provided that the signal is not too weak.

Finally we did a simulation study comparing SBL and lasso. We find that the performance of SBL depends on the strength of the signal (defined here as the minimum of the absolute value of the non-zero coefficients). With a weak signal SBL performs poorly compared to lasso, but outperforms lasso when the signal is strong.

The paper is organized as follows. We introduce the SBL method at the beginning of Section 2. We study the computation and the sparsity structure of the SBL estimator in Sections 2.1-2.3. The hard-thresholded estimator is defined and studied in Section 2.2. The simulation study is reported in Section 2.4, and all the technical proofs are grouped in Section 4. We end the paper with some open problems in Section 3.

2. Sparse Bayesian learning of linear regression models

Suppose that we observe a vector $y \in \mathbb{R}^n$ that is a realization of a random variable Y such that

$$Y = X\beta_\star + \epsilon, \tag{1}$$

for a known and non-random design matrix $X \in \mathbb{R}^{n \times p}$, a vector $\beta_\star \in \mathbb{R}^p$, and a random error term $\epsilon \in \mathbb{R}^n$ such that

$$\mathbb{E}(\epsilon) = 0, \quad \text{and} \quad \mathbb{E}(\epsilon \epsilon') = \sigma_\star^2 \mathbb{I}_n, \tag{2}$$

for $\sigma_\star^2 > 0$, where $\mathbb{I}_n$ is the n-dimensional identity matrix. Our objective is to estimate $\beta_\star$ and $\sigma_\star^2$. Although (1-2) does not make any specific distributional assumption on Y, we will consider the following possibly misspecified model: $Y \sim \mathbf{N}(X\beta, \sigma^2 I_n)$, with parameter $(\beta, \sigma^2) \in \mathbb{R}^p \times (0, \infty)$, where $\mathbf{N}(\mu, \Sigma)$ denotes the Gaussian distribution with mean μ and covariance matrix Σ. The parameter σ^2 is taken as fixed, and we assign to β a prior distribution of the form

$$\pi_\gamma(\mathrm{d}\beta) \stackrel{\text{def}}{=} \prod_{j=1}^{p} p_{\gamma_j}(\mathrm{d}\beta_j). \tag{3}$$

for a (hyper)-parameter $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Theta \stackrel{\text{def}}{=} [0, \infty)^p$, where for $a > 0$, p_a denotes the distribution of $\mathbf{N}(0, a)$, the Gaussian distribution on $\mathbb{R}$ with mean 0 and variance a, and $p_0(\mathrm{d}u) \stackrel{\text{def}}{=} \delta_0(\mathrm{d}u)$ denotes the Dirac measure at 0. The posterior distribution of β given $Y = y$ and given the hyper-parameter (γ, σ^2) is therefore

$$\pi_n(\mathrm{d}\beta | y, \sigma^2, \gamma) \propto \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left(-\frac{1}{2\sigma^2} \|y - X\beta\|^2 \right) \pi_\gamma(\mathrm{d}\beta). \tag{4}$$

Sampling from the posterior distribution $\pi_n(\cdot | y, \sigma^2, \gamma)$ is straightforward. Indeed, for $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Theta$, denote $\mathsf{I}_\gamma \stackrel{\text{def}}{=} \{1 \leq j \leq p : \gamma_j \neq 0\}$ the sparsity structure defined by γ. Notice that for $j \notin \mathsf{I}_\gamma$ (that is $\gamma_j = 0$), π_γ puts probability mass 1 on the event $\{\beta_j = 0\}$, and so does the posterior distribution $\pi_n(\cdot | y, \sigma^2, \gamma)$. Hence $\pi_n(\cdot | y, \sigma^2, \gamma)$ is the distribution of the random variable $(B_1, \ldots, B_p)$ obtained by simulating $\{B_j, j \in \mathsf{I}_\gamma\}$ from $\mathbf{N}(\mu_\gamma, \sigma^2 V_\gamma)$, and by setting the remaining components to 0, where

$$\mu_\gamma = V_\gamma X'_\gamma y, \quad V_\gamma = \left(X'_\gamma X_\gamma + \sigma^2 \bar{\Gamma}_\gamma^{-1} \right)^{-1}, \tag{5}$$

where X_γ is the matrix obtained from X by removing the columns j for which $\gamma_j = 0$, and $\bar{\Gamma}_\gamma$ is the diagonal matrix with diagonal elements given by $\{\gamma_j, j \in \mathsf{I}_\gamma\}$. With this Gaussian linear model, and prior (3), it is easy to check that the marginal distribution of y is $\mathbf{N}(0, C_\gamma)$, where

$$C_\gamma \stackrel{\text{def}}{=} \sigma^2 \mathbb{I}_n + \sum_{j \in \mathsf{I}_\gamma} \gamma_j x_j x'_j,$$

and x_j is the j-th column of X. Therefore, up to a normalizing constant that we ignore, the log-likelihood of (σ^2, γ) is given by

$$\ell(\sigma^2, \gamma) \stackrel{\text{def}}{=} -\frac{1}{2} \log \det(C_\gamma) - \frac{1}{2} \mathsf{Tr}\left(C_\gamma^{-1} yy'\right).$$

The sparse Bayesian learning (SBL) estimator of $\beta_\star$ as proposed by Tipping (2001); Faul and Tipping (2002) is the empirical Bayes estimator of β given by

$$\hat{\beta}_n = \int \beta \pi_n(\mathrm{d}\beta | y, \hat{\sigma}_n^2, \hat{\gamma}_n), \tag{6}$$

where

$$(\hat{\sigma}_n^2, \hat{\gamma}_n) = \mathsf{Argmax}_{(\sigma^2, \gamma) \in \mathbb{R}_+ \times \Theta} \, \ell(\sigma^2, \gamma). \tag{7}$$

Notice that $\hat{\beta}_n$ is straightforward to compute once $\hat{\sigma}_n^2$ and $\hat{\gamma}_n$ are available. Indeed given $\hat{\sigma}_n^2$ and $\hat{\gamma}_n$, $\hat{\beta}_{n,j} = 0$ for all j such that $\hat{\gamma}_{n,j} = 0$, and for the other components $j \in \mathsf{I}_{\hat{\gamma}_n}$, we have from (5) that

$$(\hat{\beta}_{n,j})_{j \in \mathsf{I}_{\hat{\gamma}_n}} = \left(X'_{\hat{\gamma}_n} X_{\hat{\gamma}_n} + \hat{\sigma}_n^2 \bar{\Gamma}_{\hat{\gamma}_n}^{-1}\right)^{-1} X'_{\hat{\gamma}_n} y.$$

Remark 1. The presentation of the SBL given above is slightly different from the original presentation of Tipping (2001); Faul and Tipping (2002). The key difference here is that in the prior distribution π_γ we allow the components of γ to take the value zero. This is needed for the estimator $\hat{\gamma}_n$ to be well-defined, and for the well-posedness of the question of whether the procedure produces sparse solutions.

$\square$

Computationally, the optimization problem (7) is not a "nice" problem because the objective function $\ell(\sigma^2, \gamma)$ is non-concave and typically attains its maximum at the boundary of the domain Θ (that is some of the components of its solution(s) are exactly zeros). We return to the issue of solving (7) in Section 2.3. But statistically (7) is interesting as it yields a sparse solution $\hat{\gamma}_n$ as we shall see.

2.1. **Existence of $\hat{\gamma}_n$.** Since the log-likelihood function ℓ is not concave in general, it is not immediately clear that the optimization problem (7) has a solution. It is even less clear whether the solution is sparse. Focusing on the case where σ^2 is assumed known, we show that a solution always exists.

Proposition 2. *Fix $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\sigma^2 = \sigma_\star^2$. Then the maximization problem $\mathsf{Argmax}_{\gamma \in \Theta} \ell(\gamma, \sigma^2)$ has at least one solution $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ which has the*

following property:

$$\hat{\gamma}_j = \begin{cases} \frac{(x_j'C_j^{-1}y)^2 - x_j'C_j^{-1}x_j}{(x_j'C_j^{-1}x_j)^2} & if \quad \left(x_j'C_j^{-1}y\right)^2 > x_j'C_j^{-1}x_j \\ 0 & otherwise \,, \end{cases} \tag{8}$$

where C_j is given by

$$C_j \overset{\text{def}}{=} \sigma^2\mathbb{I}_n + \sum_{k \in I_{\hat{\gamma}}\setminus\{j\}} \hat{\gamma}_k x_k x_k'.$$

Proof. See Section 4.1. □

It is important to notice that there is no randomness involved in the above result: y and X are given and fixed. In particular we do not assume (1) nor (2). It is clear that this result does not give the expression of the maximizer since the right-hand side of (8) also depends on $\hat{\gamma}$. Rather it gives coherence relationships between components of the solution. But more importantly the proposition shows that the optimization problem (7) leads to sparse solutions $\hat{\gamma}$. One can interpret the term $x_j'C_j^{-1}y$ as a measure of correlation between the y and the j-th column x_j of X. Hence the result shows that if the correlation between x_j and y is sufficiently weak then $\hat{\gamma}_{n,j}$ (and hence $\hat{\beta}_{n,j}$) is set exactly equal to zero. Of course Proposition 2 is useful only to the extent that the inequality $\left(x_j'C_j^{-1}y\right)^2 \le x_j'C_j^{-1}x_j$ is satisfied with high probability when $\beta_{\star,j} = 0$. We investigate this below. Unfortunately we will see that in general $\hat{\gamma}$ does not recover exactly the sparsity structure of $\beta_\star$, even in the most favorable setting. We make the following distributional assumption.

H1. *The data generating model (1-2) holds and $\epsilon \sim \boldsymbol{N}(0, \sigma_\star^2 I_n)$, for some $\sigma_\star^2 > 0$.*

We shall also focus our analysis on the idealized case where the matrix X has orthogonal columns.

H2. *The design matrix $X \in \mathbb{R}^{n \times p}$ is such that $\langle x_k, x_j \rangle = 0$ whenever $j \ne k$.*

Proposition 3. *Suppose that H1-2 hold, and $\sigma^2 = \sigma_\star^2$. Then for any $j \in \{1, \dots, p\}$ such that $\beta_{\star,j} = 0$,*

$$\mathbb{P}\left[\hat{\gamma}_{n,j} = 0\right] = \mathbb{P}\left[Z^2 \le 1\right] \approx 0.68,$$

where $Z \sim \boldsymbol{N}(0, 1)$.

Proof. See Section 4.2. □

The result above shows that even in the idealized setting of H2, and under the Gaussian linear model assumption, the SBL procedure will set $\hat{\gamma}_j$ to 0 (for $j \notin I$) only about 70% of the time, regardless of the sample size. We do not know whether this result continue to hold for more general design matrices. The behavior of the solution of (7) for a general design matrix is technically more challenging.

Another important limitation of the SBL procedure is the computation of $\hat{\gamma}_n$ and $\hat{\sigma}_n^2$. Typically iterative methods (such as the EM algorithm, see Section 2.3) are used. The EM algorithm does not promote sparsity, and converges to the solution only at the limit. Therefore, in finite time, the solutions generated by the EM algorithm are typically not sparse at all.

These two shortcomings limit the usefulness of the basic SBL procedure as an interesting method for sparse signal recovery. However, we observe that when $\beta_{\star,j} = 0$, and the condition $\left(x_j' C_{j,\hat{\gamma}_n}^{-1} Y \right)^2 \leq x_j' C_{j,\hat{\gamma}_n}^{-1} x_j$ fails, assuming again the most favorable setting of H2, $\hat{\gamma}_j$ is given

$$\hat{\gamma}_{n,j} = \frac{\sigma^2 (Z_j^2 - 1)}{\langle x_j, x_j \rangle},$$

where $Z_j \sim \mathbf{N}(0,1)$. Hence $\hat{\gamma}_j$ has mean zero and variance of order $O(\|x_j\|^{-4}) \approx O(n^{-2})$. We conclude that when SBL fails to set to zero a component j such that $\beta_{\star,j} = 0$, the computed SBL solution $\hat{\gamma}_j$ is typically very small. This suggests that a thresholded version of $\hat{\gamma}_n$ should be able to set these terms to zero. We pursue this approach in Section 2.2.

2.2. A thresholded version and its statistical properties.
We saw in Section 2.1 that although sparse, $\hat{\gamma}_n$ does not recover in general the sparsity structure of $\beta_\star$. To improve on this we propose a modified, hard-thresholded version of $\hat{\gamma}_n$ denoted $\tilde{\gamma}_n$ and defined as follows. For $1 \leq j \leq p$,

$$\tilde{\gamma}_{n,j} \overset{\text{def}}{=} \begin{cases} \hat{\gamma}_{n,j} & \text{if } \hat{\gamma}_{n,j} > \frac{\hat{\sigma}_n^2 z_\star}{\|x_j\|^2} \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

for a thresholding parameter $z_\star$ that we set to $z_\star = c(1 + |\hat{\rho}|) \log p$, for a constant c, and where $\hat{\rho}$ is an estimate of the largest correlation among the columns of X. The corresponding modified estimator of $\beta_\star$ is

$$\tilde{\beta}_n \overset{\text{def}}{=} \int \beta \pi_n(\mathrm{d}\beta | y, \hat{\sigma}_n^2, \tilde{\gamma}_n).$$

Theorem 4. *Assume H1-2, and suppose that $\sigma_\star^2$ is known, $\log s \geq 1$, and $z_\star = c_0 \log p$, for some constant $c_0 > 2$, where $s = |I_{\gamma_\star}|$. Then*

$$\|\tilde{\beta}_n - \beta_\star\|_2^2 \leq M \frac{\sigma^2 s \log(p)}{n}, \tag{10}$$

with probability at least $1 - \frac{1}{p^{(c_0 s)/8}} - \frac{1}{\exp(s)}$, where $M = \frac{4(2+c_0)}{c}$, and $c = \min_{1 \leq i \leq p} \|x_j\|^2/n$.

Proof. See Section 4.3. $\qquad\qquad\square$

We deduce the following corollary. For $u \in \mathbb{R}^p$, $\mathsf{sign}(u) = (s_1, \ldots, s_p)$ where for each i, $s_i = 0$ if $u_i = 0$, $s_i = 1$ is $u_i > 0$, and $s_i = -1$ if $u_i < 0$.

Corollary 5. *In addition to the assumptions of Theorem 4, suppose that*

$$\min_{\{j:\, |\beta_{\star,j}|>0\}} |\beta_{\star,j}| > \sqrt{\frac{M\sigma^2 s \log p}{n}}. \tag{11}$$

Then with probability at least $1 - \frac{1}{p^{\frac{c_0 s}{8}}} - \frac{1}{\exp(s)} - \frac{1}{p^{\frac{c_0}{2}-1}}$, $\mathsf{sign}(\tilde{\beta}_n) = \mathsf{sign}(\beta_\star)$.

Proof. See Section 4.4. $\qquad\qquad\square$

2.3. **Computing $\hat{\sigma}_n^2$ and $\hat{\gamma}_n$.** Here we address the issue of solving (7). Because the function $\ell(\sigma^2, \gamma)$ is not concave, and typically attains its maximum at the boundary of the domain Θ, the optimization (7) is not a smooth problem. The strategy originally developed by Tipping (2001) focuses instead on the smooth problem obtained by maximizing ℓ over the open domain $\mathbb{R}_+^{p+1}$, where $\mathbb{R}_+ \overset{\text{def}}{=} (0, \infty)$. That is, find

$$\mathsf{Argmax}_{(\sigma^2, \gamma) \in \mathbb{R}_+^{p+1}} \ell(\sigma^2, \gamma). \tag{12}$$

Of course, this latter problem has no solution whenever the solution of (7) occurs at the boundary of Θ. Nevertheless, we will see that an EM algorithm that attempts to solve (12) produces sequences that converge to the solution of (7).

Since the likelihood function $\exp(\ell)$ of (σ^2, γ) is obtained by integrating out β, we can treat β as a missing variable and use the EM algorithm as proposed by Tipping (2001). For $\gamma \in \mathbb{R}_+^p$, the so-called complete log-likelihood takes the form

$$\ell_{\mathsf{com}}(\beta, \sigma^2, \gamma|y) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|y - X\beta\|^2 - \frac{1}{2} \sum_{j=1}^p \left(\log \gamma_j + \frac{\beta_j^2}{\gamma_j} \right).$$

Given a working solution $(\{\sigma^2\}^{(k)}, \gamma^{(k)}) \in \mathbb{R}_+^{p+1} = (0, \infty)^{p+1}$, set

$$Q(\sigma^2, \gamma | \{\sigma^2\}^{(k)}, \gamma^{(k)}) \overset{\text{def}}{=} \int \ell_{\mathsf{com}}(\beta, \sigma^2, \gamma|y) \pi_n(d\beta|y, \{\sigma^2\}^{(k)}, \gamma^{(k)}),$$

the so-called Q-function. We will use the upper-script (k) to index sequences generated by the EM algorithm. Set $V^{(k)} = \left(X'X + \{\sigma^2\}^{(k)}\{\Gamma^{(k)}\}^{-1}\right)^{-1}$, and $\mu^{(k)} = V^{(k)}X'y$, where $\Gamma^{(k)} = \mathsf{diag}(\gamma_1^{(k)}, \ldots, \gamma_p^{(k)})$. Maximizing $Q(\cdot|\{\sigma^2\}^{(k)}, \gamma^{(k)})$ is easy and gives

$$
\begin{aligned}
\{\sigma^2\}^{(k+1)} &= \frac{1}{n} \int \|y - X\beta\|^2 \pi_n(\beta|\{\sigma^2\}^{(k)}, \gamma^{(k)}, y)\mathrm{d}\beta \\
&= n^{-1}\left(\|y - X\mu^{(k)}\|^2 + \{\sigma^2\}^{(k)}\mathsf{Tr}(V^{(k)}X'X)\right),
\end{aligned}
$$

and for $j = 1, \ldots, p$,

$$
\gamma_j^{(k+1)} = \{\mu_j^{(k)}\}^2 + \{\sigma^2\}^{(k)}V_{j,j}^{(k)}.
$$

This leads to the following algorithm for solving (12)

Algorithm 1 (EM algorithm). Given $(\{\sigma^2\}^{(k)}, \gamma^{(k)}) \in \mathbb{R}_+^{p+1} = (0, \infty)^{p+1}$, we compose the matrix $\Gamma^{(k)} = \mathsf{diag}(\gamma_1^{(k)}, \ldots, \gamma_p^{(k)})$.

(1) Compute $V^{(k)} = \left(X'X + \{\sigma^2\}^{(k)}\{\Gamma^{(k)}\}^{-1}\right)^{-1}$, and $\mu^{(k)} = V^{(k)}X'y$.
(2) Set

$$
\gamma_j^{(k+1)} = \{\mu_j^{(k)}\}^2 + \{\sigma^2\}^{(k)}V_{j,j}^{(k)}, \quad j = 1, \ldots, p,
$$
$$
\{\sigma^2\}^{(k+1)} = \frac{1}{n}\left(\|y - X\mu^{(k)}\|^2 + \{\sigma^2\}^{(k)}\mathsf{Tr}(V^{(k)}X'X)\right).
$$

Although this EM algorithm is designed to solve the maximization problem (12) we will see that it typically converges to the solution of (7). To simplify the analysis we assume again that H2 holds and that σ^2 is fixed. Hence we focus only on the recursion in γ:

$$
\gamma_j^{(k+1)} = \{\mu_j^{(k)}\}^2 + \sigma^2 V_{j,j}^{(k)}, \quad j = 1, \ldots, p.
$$

With the assumption that the design matrix is orthogonal, we can work out explicitly the terms $V^{(k)} = \left(X'X + \sigma^2\{\Gamma^{(k)}\}^{-1}\right)^{-1}$ and $\mu^{(k)} = V^{(k)}X'y$, which leads to

$$
\gamma_j^{(k+1)} = \frac{\langle x_j, y\rangle^2}{\left(\|x_j\|^2 + \frac{\sigma^2}{\gamma_j^{(k)}}\right)^2} + \frac{\sigma^2}{\|x_j\|^2 + \frac{\sigma^2}{\gamma_{j(k)}}}, \quad j = 1, \ldots, p. \tag{13}
$$

Proposition 6. *Fix $y \in \mathbb{R}^n$, and $X \in \mathbb{R}^{n \times p}$ such that H2 holds. Fix $\sigma^2 > 0$. Let $\{\gamma^{(k)}, k \geq 0\}$ denote the sequence produced by the recursion (13) for some initial $\gamma^{(0)}$ with positive components. Then for all $j \in \{1, \ldots, p\}$,*

$$
\lim_{k \to \infty} \gamma_j^{(k)} = \hat{\gamma}_{n,j} = \begin{cases} \frac{\langle y, x_j\rangle^2 - \sigma^2\|x_j\|^2}{\|x_j\|^2} & \text{if } \langle y, x_j\rangle^2 > \sigma^2\|x_j\|^2 \\ 0 & \text{otherwise} \end{cases}.
$$

Proof. See Section 4.5   □

Remark 7. In the non-orthogonal design setting, our simulation results suggest that the conclusion of Proposition 2 continues to hold, although we do not have any rigorous proof.

2.4. A simulation study.

2.4.1. *Synthetic Data Sets.* We investigate by simulation the behavior of the SBL procedure and its thresholding version, and how they compare with lasso. For all the simulations, $n = 100$ and $p = 500$. We generate the design matrix X by simulating each row independently from the Gaussian distribution $\mathbf{N}(0, \Sigma)$ where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$. We consider two values of ρ: $\rho = 0$ for which X is close to satisfy H2, and $\rho = 0.9$ which produces a design matrix X with strongly correlated variables. We simulate the dependent variable Y from the $\mathbf{N}(X\beta_\star, \sigma_\star^2 \mathbb{I}_n)$, with $\sigma_\star = 1$. We consider four (4) different scenarios of sparsity, with $s = 3, 15, 25$, and $s = 50$ where s is the number of non-zero elements of $\beta_\star$. The magnitude of the non-zero elements also play an important role in the recovery. We generate all the non-zeros components of $\beta_\star$ from the uniform distribution $\mathbf{U}(a, a+1)$, for a ranging from 0 to 9.

For each value of ρ, each sparsity level, and each signal strength a, we repeat each estimator 30 times, and we compute the relative error rate ($\|\hat{\beta} - \beta_\star\| / \|\beta_\star\|$), the sensitivity and the specificity, averaged over these 30 replications. The sensitivity (SEN) and the specificity (SPE) of a given estimator $\hat{\beta}$ are defined as

$$\mathsf{SEN}(\hat{\beta}) = \frac{\sum_{j=1}^{p} \mathbf{1}_{\{\hat{\beta}_j \neq 0\}} \mathbf{1}_{\{\beta_{\star,j} \neq 0\}}}{\sum_{j=1}^{p} \mathbf{1}_{\{\beta_{\star,j} \neq 0\}}}, \quad \text{and} \quad \mathsf{SPE}(\hat{\beta}) = \frac{\sum_{j=1}^{p} \mathbf{1}_{\{\hat{\beta}_j \neq 0\}} \mathbf{1}_{\{\beta_{\star,j} \neq 0\}}}{\sum_{j=1}^{p} \mathbf{1}_{\{\hat{\beta}_j \neq 0\}}}.$$

These measures are valid for any estimator $\hat{\beta}$, and we compute them for the thresholded version of SBL, the non-thresholded version of SBL, as well as for the lasso estimator. For the thresholded SBL, we use $z_\star = c(1 + |\hat{\rho}|) \log p$, where c is determined by minimizing the BIC: $\frac{\|y - X\hat{\beta}\|}{2\hat{\sigma}^2} + s \log(n)$.

We compute the lasso estimator using the function *cv.glmnet* of the package GLM-Net (Friedman et al. (2010)) where we select the penalty term λ by a 10-fold cross-validation procedure. In the cross-validation, the regulation parameter selected minimizes the prediction error.

The simulation results are presented on Figure 1-8. As one can see from these figures, the main conclusion is that SBL is more sensitive than lasso to the strength of the signal (defined here as as the parameter a). With a weak signal it performs

poorly, but outperforms lasso when the signal is strong enough. Another interesting finding is that, overall, lasso performs poorly in selecting the non-zeros components (variable selection). This is consistent with recent results (Meinhausen and Yu (2009)) which shows that variable selection consistency of lasso requires the irrepresentable condition, which actually is a very strong condition that often does not hold in practice. For instance, the irrepresentable condition fails for all the design matrices of this simulation study, except for the design matrix behind Figure 2.

2.4.2. *A simulated real data example.* In this example, we consider a micro-array data concerning genes involved in the production of riboflavin. The data is made publicly available at

```
http://www.annualreviews.org/doi/suppl
                    /10.1146/annurev-statistics-022513-115545
```

and contains $n = 71$ samples and $p = 4088$ covariates corresponding to 4088 genes. Each of the sample contains a real valued response consisting of the logarithm of the riboflavin production rate and 4088 real valued covariates consisting of the logarithm of the genes' expression levels.

Given the very high dimensionality of this dataset, the lack of any true value of the parameter, and given also the fact that micro-array data are well-known to be very noisy, direct comparison of different regression methods on such dataset cannot be very insightful. For a more meaningful comparison, we use the riboflavin design matrix $X \in \mathbb{R}^{71 \times 4088}$ to generate simulated levels of riboflavin production rate using the sparse regression model $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 \mathbb{I}_{71})$, with $\sigma^2 = 1$. The magnitude of the non-zero components of β are uniformly simulated $\beta_j \sim U(a, a+1)$ with $a = \{0, 1, \ldots, 9\}$. We set the number of non-zeros elements in the vector β to 5. Figure 9 shows the results of the simulation evaluated using the aforementioned metrics. Under such extreme high-dimensional conditions, both methods perform poorly. SBL has found all the relevant variables but has also selected many non-relevant variables. Lasso has produced more sparse solutions, but has missed some important variables. The results remain essentially the same even when we set σ^2 (the variance of the noise ϵ) to 0.1.

One final word on computing times. We compute the SBL estimate using Algorithm 1, and we use the package GLMNet to compute lasso. We implemented Algorithm 1 in R. The core of the GLMNet package is written in Fortran and the result is very fast.

The comparison of the computing times is largely in favor of GLMNet. Comparing computing times is always tricky as it depends to a large extent on the programming language and skills. But beyond the implementation differences, it seems clear that lasso has a computational advantage over SBL in that it leads to "easier" (convex) optimization problems, compared to SBL.

3. Conclusion

We have shown that when the design matrix is orthogonal, the SBL estimator is uniquely defined, sparse (however does not recover the true sparsity structure of the signal), and can be computed using the EM algorithm. We have also proposed a hard-thresholded version of SBL, and shown that the hard-thresholded estimator recovers the true sparsity structure of the model, and achieves the same estimation error bound as lasso (with high probability). Furthermore our simulation results show that the method compares very well with lasso, and outperforms lasso when the regression coefficients are not too small.

One important and pressing issue is the extension of these results to non-orthogonal design matrices. In particular we wish to understand the type of design matrix X for which these results continue to hold. This SBL theory and its comparison with the recently developed lasso theory (see for instance Meinhausen and Yu (2009); Bickel et al. (2009)) could potentially give new insight into high-dimensional regression analysis. The generalized singular value decomposition (see e.g. Golub and Van Loan (2013)) of X_γ and Γ_γ seems to be a promising approach to tackle this problem. The challenge in this approach appears to be the development of an appropriate differentiability theory for the components of the GSVD decomposition as a function of γ.

The SBL method can be extended in several directions. It can be easily extended to deal with generalized linear models, and graphical models. But in these extensions, the computation of the estimator might require some new algorithms. Another possible extension of the method would be to replace the Gaussian distribution in the prior π_γ by some other distribution. Some of these extensions of the methodology are already being explored. For instance Balakrishnan and Madigan (2010) replaced the Gaussian distribution by the double-exponential distribution and shows by simulation that the resulting estimator (called demi-lasso) compares very well with lasso and the standard SBL.
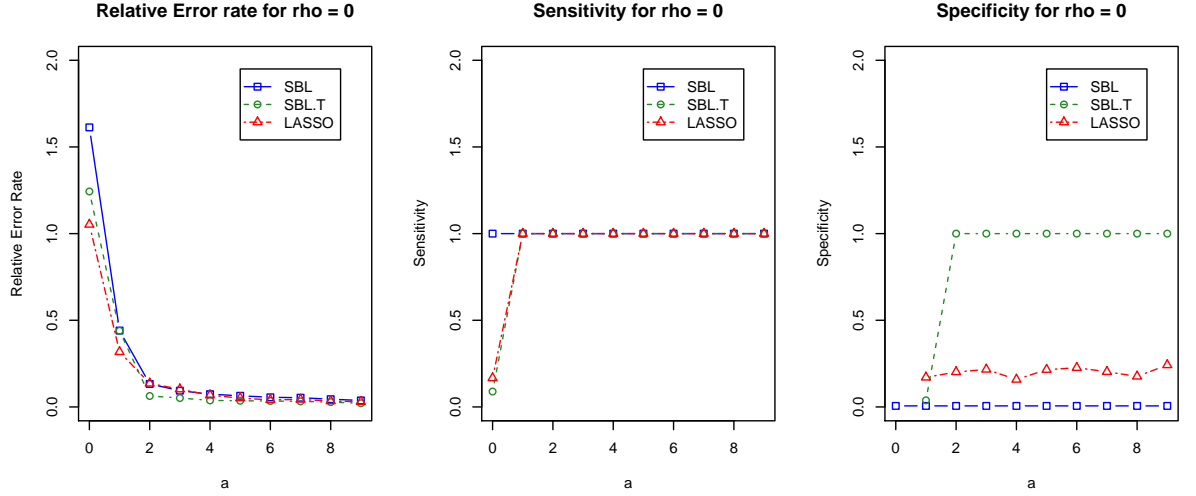
Figure 1: Sensitivity, specificity and relative error for SBL and lasso as function of
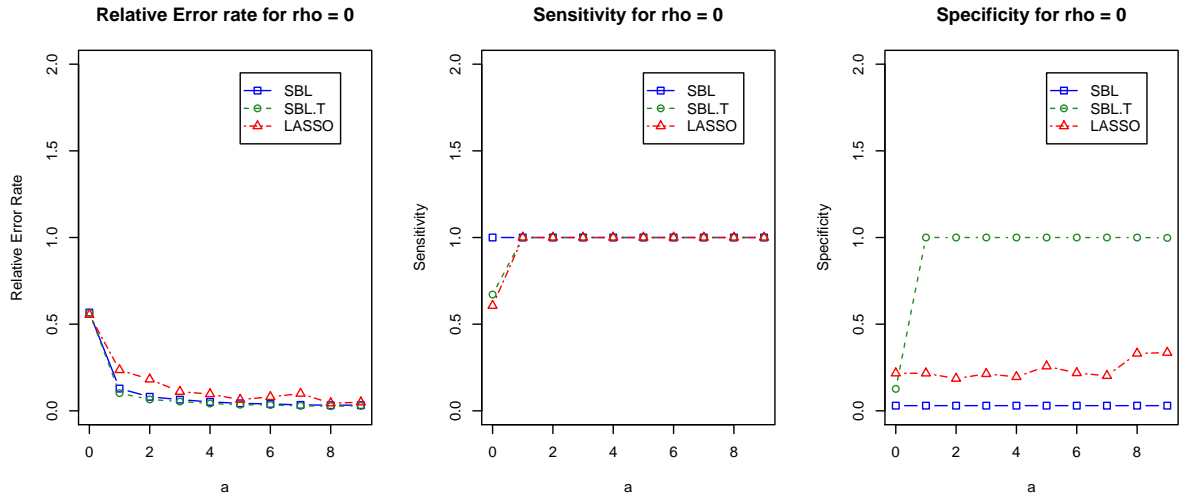a. $s = 3$.

Figure 2: Sensitivity, specificity and relative error for SBL and lasso as function of
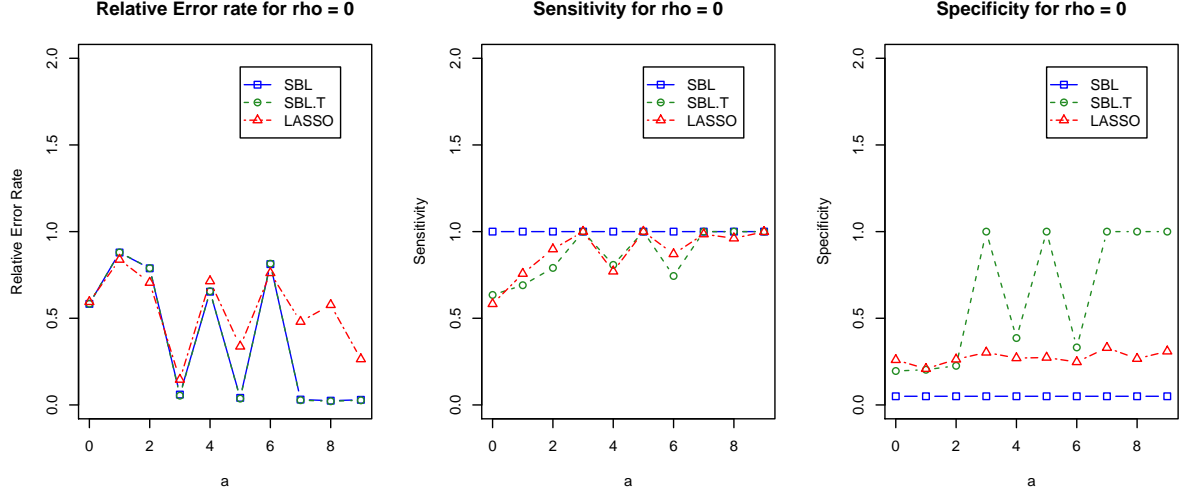a. $s = 15$.

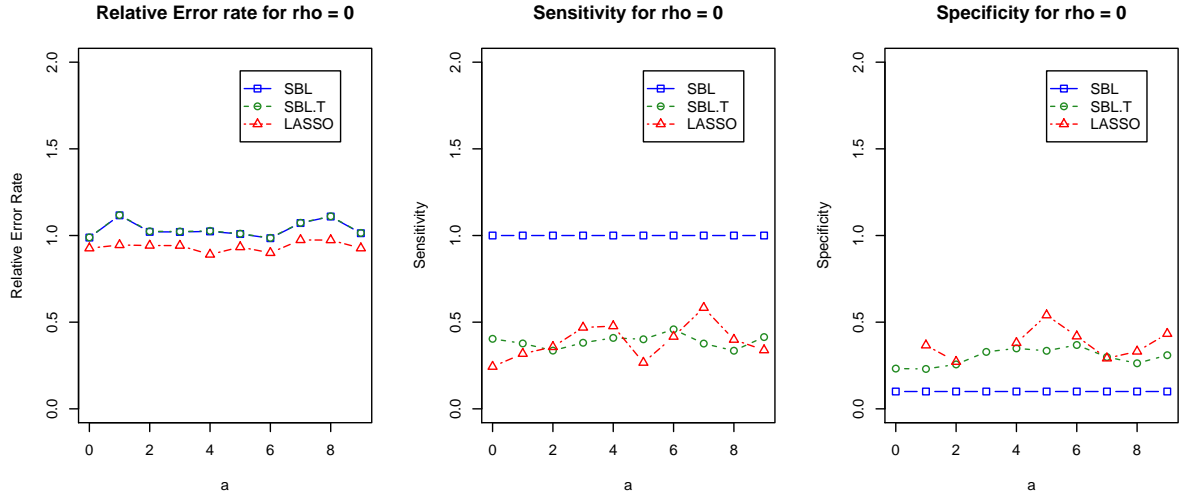Figure 3: Sensitivity, specificity and relative error for SBL and lasso as function of a. $s = 25$.

Figure 4: Sensitivity, specificity and relative error for SBL and lasso as function of a. $s = 50$.
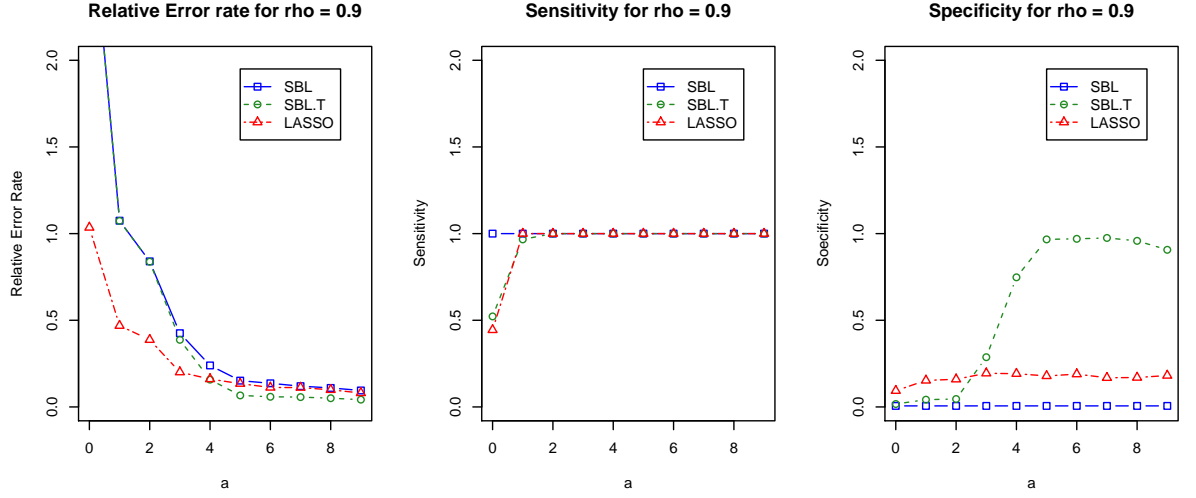
Figure 5: Sensitivity, specificity and relative error for SBL and lasso as function of a. $s = 3$.
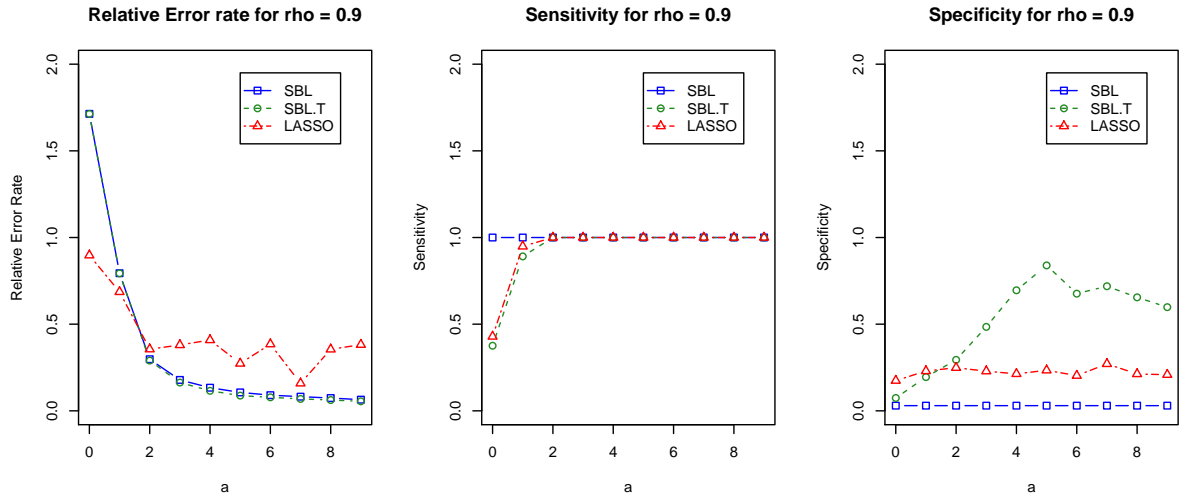

Figure 6: Sensitivity, specificity and relative error for SBL and lasso as function of a. $s = 15$.
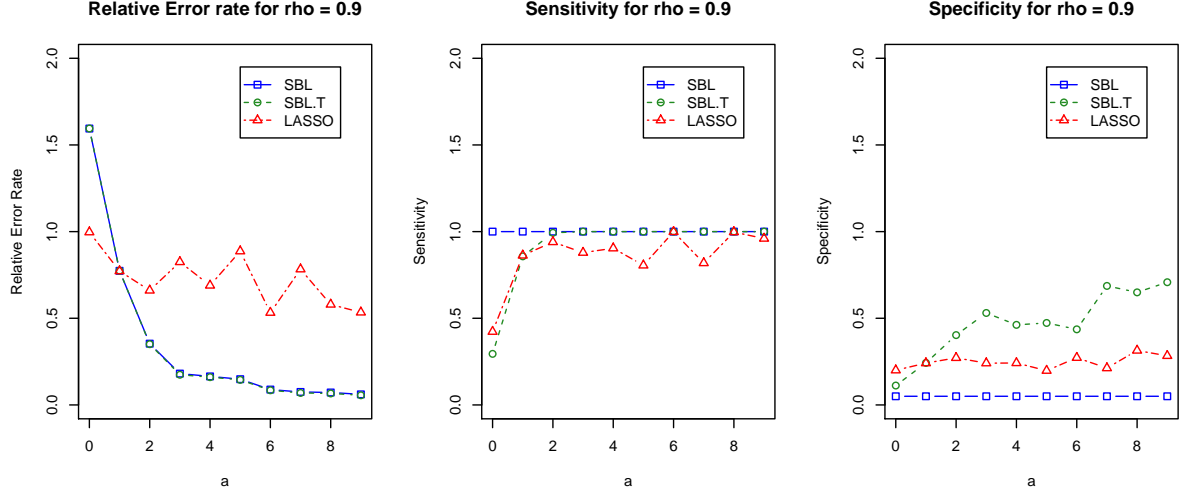
Figure 7: Sensitivity, specificity and relative error for SBL and lasso as function of a. $s = 25$.
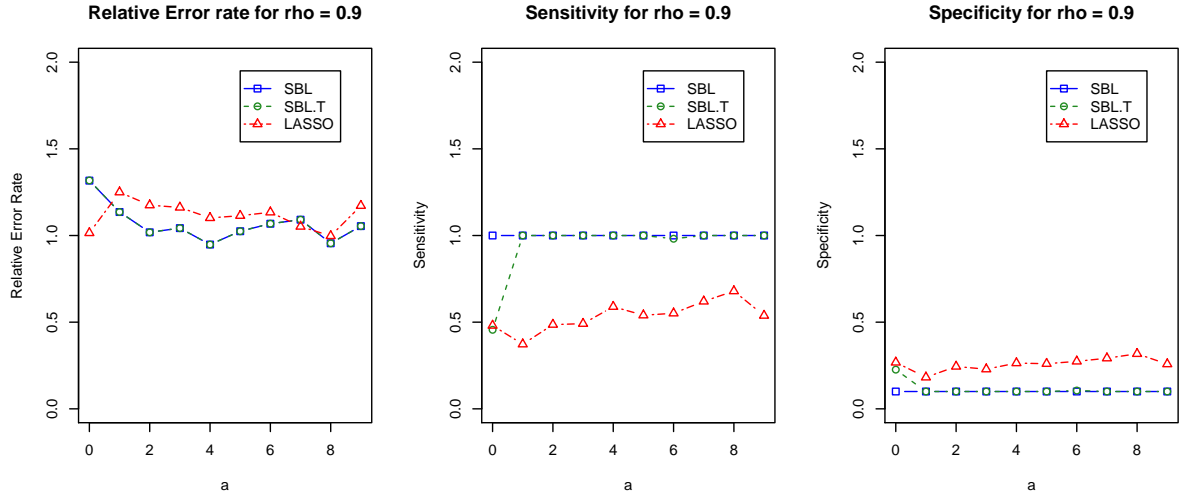

Figure 8: Sensitivity, specificity and relative error for SBL and lasso as function of a. $s = 50$.
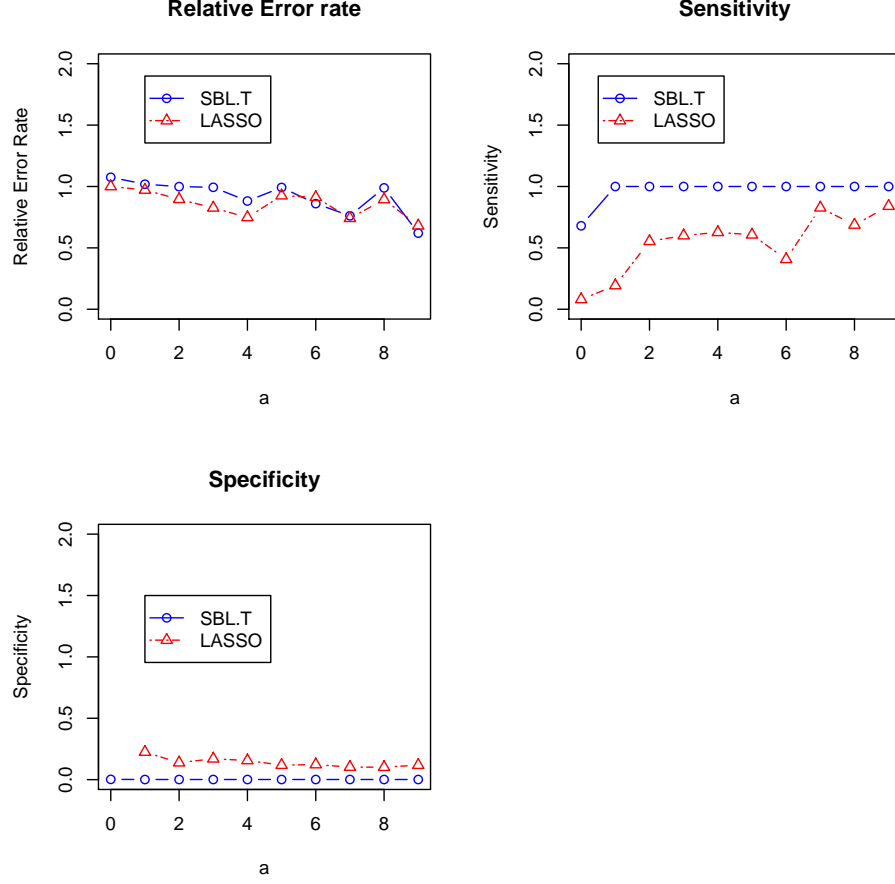
<u>Figure 9</u>: Sensitivity, specificity and relative error for SBL(thresholded) and lasso as function of a. $s = 5$. The results of this simulation is generated using the riboflavin data

4. PROOFS

4.1. **Proof of Proposition 2.**

Proof. For any $i \in \{1, \ldots, p\}$,

$$\ell(\gamma) \le -\frac{1}{2} \log \det(C_\gamma) \le -\frac{1}{2} \log \det \left(\sigma^2 \mathbb{I}_n + \gamma_i x_i x_i' \right) \downarrow -\infty,$$

as $\gamma_i \to \infty$. This together with the continuity of $\gamma \mapsto \ell(\gamma)$ imply the existence of a maximizer. For any such maximizer $\hat{\gamma}$, consider the vector γ such that the j-th component of γ is free to vary and the remaining components γ_{-j} are fixed to $\hat{\gamma}_{-j}$. Then we write $C_\gamma = C_j + \gamma_j x_j x_j'$ and use the matrix identity $(A + uu')^{-1} =$

$A^{-1} - \frac{A^{-1}uu'A^{-1}}{1+u'A^{-1}u}$ to deduce that

$$C_\gamma^{-1} = C_j^{-1} - \frac{\gamma_j C_j^{-1} x_j x_j' C_j^{-1}}{1 + \gamma_j x_j' C_j^{-1} x_j}.$$

Therefore,

$$\ell(\gamma) = -\frac{1}{2} \log \det \left(C_j + \gamma_j x_j x_j' \right) + \frac{1}{2} \frac{\gamma_j \left(x_j' C_j^{-1} y \right)^2}{1 + \gamma_j x_j' C_j^{-1} x_j} - \frac{1}{2} y' C_j^{-1} y.$$

Since C_j does not depend on γ_j, we easily see that $\gamma_j \mapsto \ell(\gamma)$ is differentiable on $(0, \infty)$ and

$$
\begin{aligned}
\frac{\partial}{\partial \gamma_j} \ell(\gamma) &= -\frac{1}{2} x_j' C_j^{-1} x_j + \frac{1}{2} \frac{\gamma_j \left(x_j' C_j^{-1} x_j \right)^2}{1 + \gamma_j x_j' C_j^{-1} x_j} + \frac{1}{2} \frac{\left(x_j' C_j^{-1} y \right)^2}{\left(1 + \gamma_j x_j' C_j^{-1} x_j \right)^2} \\
&= \frac{1}{2} \frac{\left(x_j' C_j^{-1} y \right)^2 - \gamma_j \left(x_j' C_j^{-1} x_j \right)^2 - x_j' C_j^{-1} x_j}{\left(1 + \gamma_j x_j' C_j^{-1} x_j \right)^2}
\end{aligned}
$$

If $\left(x_j' C_j^{-1} y \right)^2 \leq x_j' C_j^{-1} x_j$, $\frac{\partial}{\partial \gamma_j} \ell(\gamma) < 0$ and $\gamma_j \mapsto \ell(\gamma)$ attains its maximum at 0. Similarly if $\left(x_j' C_j^{-1} y \right)^2 > x_j' C_j^{-1} x_j$, it is easy to check that $\gamma_j \mapsto \ell(\gamma)$ attains its maximum at $\left(\left(x_j' C_j^{-1} y \right)^2 - x_j' C_j^{-1} x_j \right) / \left(x_j' C_j^{-1} x_j \right)^2$. Now if $\hat{\gamma}_j$ differs from the maximizer just found, we can improve on the likelihood by setting $\hat{\gamma}_j$ equal to that maximizer, which would be a contradiction. Hence the result.

$\square$

4.2. **Proof of Proposition 3.**

Proof. Recall that $\mathsf{I} = \{1 \leq j \leq p : \beta_{\star,j} \neq 0\}$ is the sparsity structure of $\beta_\star$. For $\gamma \in \Theta$, and $1 \leq j \leq p$, we define $\mathsf{I}_0 \overset{\text{def}}{=} \mathsf{I} \cap \mathsf{I}_\gamma \setminus \{j\}$, and $\mathsf{I}_1 \overset{\text{def}}{=} \mathsf{I}^c \cap \mathsf{I}_\gamma \setminus \{j\}$, where in order to keep the notation easy, we omit the dependence of I_0 and I_1 on (γ, j). We will also write X_{I_0} (resp. X_{I_1}) to denote the matrix obtained by collecting the columns of X whose indexes belong to I_0 (resp. I_1). We define

$$
\begin{aligned}
C_{j,\gamma} \overset{\text{def}}{=} & \sigma^2 \mathbb{I}_n + \sum_{k \in \mathsf{I}_\gamma \setminus \{j\}} \gamma_k x_k x_k' \\
= & \sigma^2 \mathbb{I}_n + \sum_{k \in \mathsf{I}_0} \gamma_k x_k x_k' + \sum_{k \in \mathsf{I}_1} \gamma_k x_k x_k'.
\end{aligned}
$$

By the Woodbury matrix identity and the assumption $X'_{\mathsf{I}_0} X_{\mathsf{I}_1} = 0$, we get:

$$C_{j,\gamma}^{-1} = \frac{1}{\sigma^2}\mathbb{I}_n - \frac{1}{\sigma^4}X_{\mathsf{I}_0}\left(\Gamma_{\mathsf{I}_0}^{-1} + \frac{1}{\sigma^2}X'_{\mathsf{I}_0}X_{\mathsf{I}_0}\right)^{-1}X'_{\mathsf{I}_0} - \frac{1}{\sigma^4}X_{\mathsf{I}_1}\left(\Gamma_{\mathsf{I}_1}^{-1} + \frac{1}{\sigma^2}X'_{\mathsf{I}_1}X_{\mathsf{I}_1}\right)^{-1}X'_{\mathsf{I}_1}.$$

Hence, for $k \in \mathsf{I}$, and using the fact that $j \notin \mathsf{I}$, we have

$$x'_j C_{j,\gamma}^{-1} x_k = 0, \quad\text{and}\quad x'_j C_{j,\gamma}^{-1} x_j = \frac{1}{\sigma^2}\|x_j\|^2.$$

Therefore, if $Y = X\beta_\star + \epsilon$, we get

$$x'_j C_{j,\gamma}^{-1} Y = \frac{1}{\sigma^2}\langle x_j, \epsilon\rangle \sim \mathbf{N}\left(0, \sigma^2\|x_j\|^2\right).$$

Now, the matrix C_j defined in Proposition 2 is $C_j = C_{j,\hat\gamma_n}$. Hence

$$\mathbb{P}\left[\hat\gamma_{n,j} = 0\right] = \mathbb{P}\left[\left(x'_j C_{j,\hat\gamma_n}^{-1} Y\right)^2 \le x'_j C_{j,\hat\gamma_n}^{-1} x_j\right] = \mathbb{P}\left[Z^2 \le 1\right],$$

where $Z \sim \mathbf{N}(0,1)$. Hence the result. $\qquad\square$

4.3. **Proof of Theorem 4.**

Proof. Under H2, $x_j C_{j,\hat\gamma_n}^{-1} Y = \langle x_j, Y\rangle / \sigma^2$, and $x_j C_{j,\hat\gamma_n}^{-1} x_j = \|x_j\|^2/\sigma^2$. Hence

$$\tilde\gamma_j = \begin{cases} \frac{\langle x_j, Y\rangle^2 - \sigma^2\|x_j\|^2}{\|x_j\|^2} & \text{if}\quad \langle x_j, Y\rangle^2 > \sigma^2\|x_j\|^2(1+z_\star) \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, under H1 $\hat\beta_{n,j}$ has the explicit form $\hat\beta_{n,j} = \frac{\langle Y, X_j\rangle}{\|x_j\|^2 + \frac{\sigma^2}{\hat\gamma_{n,j}}}$. It follows that

$$\tilde\beta_{n,j} = \begin{cases} \frac{\langle Y, x_j\rangle}{\|x_j\|^2 + \frac{\sigma^2}{\hat\gamma_{n,j}}} & \text{if}\quad \langle Y, x_j\rangle^2 > \sigma^2\|x_j\|^2(1+z_\star) \\ 0 & \text{otherwise .} \end{cases}$$

Again using the orthogonality assumption of X, we obtain $\langle Y, x_j\rangle = \beta_{\star,j}\|x_j\|^2 + \langle \epsilon, X_j\rangle$. We set $t_j \overset{\text{def}}{=} \langle \epsilon, X_j\rangle$. Then it follows that

$$\tilde\beta_{n,j} - \beta_{\star,j} = \begin{cases} 0 & \text{if } \beta_{\star,j} = 0, \text{ and } \left(\frac{t_j}{\|x_j\|^2}\right)^2 \le \frac{\sigma^2}{\|x_j\|^2}(1+z_\star) \\ \frac{t_j}{\|x_j\|^2 + \frac{\sigma^2}{\hat\gamma_{n,j}}} & \text{if } \beta_{\star,j} = 0, \text{ and } \left(\frac{t_j}{\|x_j\|^2}\right)^2 > \frac{\sigma^2}{\|x_j\|^2}(1+z_\star) \\ -\beta_{\star,j} & \text{if } \beta_{\star,j} \neq 0 \text{ and } \left(\frac{t_j}{\|x_j\|^2} + \beta_{\star,j}\right)^2 \le \frac{\sigma^2}{\|x_j\|^2}(1+z_\star) \\ \frac{\|x_j\|^2\beta_{\star,j} + t_j}{\|x_j\|^2 + \frac{\sigma^2}{\hat\gamma_{n,j}}} - \beta_{\star,j} & \text{if } \beta_{\star,j} \neq 0 \text{ and } \left(\frac{t_j}{\|x_j\|^2} + \beta_{\star,j}\right)^2 > \frac{\sigma^2}{\|x_j\|^2}(1+z_\star). \end{cases}$$

Suppose that $j \in I = I_{\gamma_\star}$ and $\left(\frac{t_j}{\|x_j\|^2} + \beta_{\star,j}\right)^2 \leq \frac{\sigma^2}{\|x_j\|^2}(1 + z_\star)$. Then with $Z_j \overset{\text{def}}{=} \frac{t_j}{\sigma\|x_j\|}$,

$$|\beta_{\star,j}| \leq \left|\frac{t_j}{\|x_j\|^2}\right| + \frac{\sigma}{\|x_j\|}\sqrt{1 + z_\star} = \frac{\sigma}{\|x_j\|}\left(|Z_j| + \sqrt{1 + z_\star}\right).$$

Hence for such index j,

$$\beta_{\star,j}^2 \leq \frac{2\sigma^2}{\|x_j\|^2}\left(Z_j^2 + 1 + z_\star\right). \tag{14}$$

But for $j \in I_{\gamma_\star}$, such that $\left(\frac{t_j}{\|x_j\|^2} + \beta_{\star,j}\right)^2 > \frac{\sigma^2}{\|x_j\|^2}(1 + z_\star)$, $\tilde{\gamma}_{n,j} = \left(\frac{t_j}{\|x_j\|^2} + \beta_{\star,j}\right)^2 - \frac{\sigma^2}{\|x_j\|^2}$. Using this with some easy algebra, we obtain that for such index j,

$$\frac{\|x_j\|^2\beta_{\star,j} + t_j}{\|x_j\|^2 + \frac{\sigma^2}{\tilde{\gamma}_{n,j}}} - \beta_{\star,j} = \frac{t_j}{\|x_j\|^2} - \frac{\sigma^2}{t_j + \|x_j\|^2\beta_{\star,j}}. \tag{15}$$

It follows that

$$\left|\frac{\|x_j\|^2\beta_{\star,j} + t_j}{\|x_j\|^2 + \frac{\sigma^2}{\tilde{\gamma}_{n,j}}} - \beta_{\star,j}\right| \leq \frac{\sigma}{\|x_j\|}\left(1 + |Z_j|\right). \tag{16}$$

With (14) and (16), we get

$$\sum_{j\in I_{\gamma_\star}}\left(\tilde{\beta}_{n,j} - \beta_{\star,j}\right)^2 \leq \frac{2\sigma^2}{cn}\sum_{j\in I_{\gamma_\star}}(1 + z_\star + Z_j^2),$$

where $c = \min_{1\leq i\leq p}\|x_j\|^2/n$. Set $s \overset{\text{def}}{=} |I_{\gamma_\star}|$. By Teicher (1984) Lemma 5, $\mathbb{E}(|Z_j^2 - 1|^k) \leq k!2^{k-2}$, $k > 2$. Hence by Bernstein's inequality (see e.g. van der Vaart and Wellner (1996) Lemma 2.2.11), we conclude that

$$\mathbb{P}\left[\sum_{j\in I_{\gamma_\star}}(1 + z_\star + Z_j^2) > 2(1 + z_\star)s\right] \leq \mathbb{P}\left[\sum_{j\in I_{\gamma_\star}}(Z_j^2 - 1) > z_\star s\right]$$

$$\leq \exp\left(-\frac{z_\star^2 s^2}{4(1 + z_\star s)}\right) \leq \exp\left(-\frac{z_\star s}{8}\right) \leq \frac{1}{p^{c_0 s/8}}.$$

Hence with probability at least $1 - \frac{1}{p^{c_0 s/8}}$,

$$\sum_{j\in I_{\gamma_\star}}\left(\tilde{\beta}_{n,j} - \beta_{\star,j}\right)^2 \leq \frac{4\sigma^2}{cn}(1 + z_\star)s \leq \frac{4(1 + c_0)}{c}\frac{\sigma^2 s\log p}{n}. \tag{17}$$

On the other hand, from (15), $\frac{t_j}{\|x_j\|^2 + \frac{\sigma^2}{\tilde{\gamma}_{n,j}}} = \frac{t_j}{\|x_j\|^2} - \frac{\sigma^2}{t_j}$, hence

$$\sum_{j \notin I_{\gamma_\star}} \left(\tilde{\beta}_{n,j} - \beta_{\star,j}\right)^2 = \sum_{j \notin I_{\gamma_\star}, Z_j^2 > 1 + z_\star} \frac{\sigma^2}{\|x_j\|^2} \left(Z_j - \frac{1}{Z_j}\right)^2 \leq \frac{\sigma^2}{cn} \sum_{j \notin I_{\gamma_\star}} Z_j^2 \mathbf{1}_{\{Z_j^2 > 1 + z_\star\}}$$

$$\leq \frac{\sigma^2}{cn} \sum_{j=1}^{p} Z_j^2 \mathbf{1}_{\{Z_j^2 > 1 + z_\star\}}.$$

Now for any $\kappa \in (0, 1/2)$, $a > 0$, and by Markov's inequality

$$\mathbb{P}\left[\sum_{j=1}^{p} Z_j^2 \mathbf{1}_{\{Z_j^2 > 1 + z_\star\}} > a\right] = \mathbb{P}\left[\exp\left(\sum_{j=1}^{p} \kappa Z_j^2 \mathbf{1}_{\{Z_j^2 > 1 + z_\star\}}\right) > e^{a\kappa}\right]$$

$$\leq \exp\left[-a\kappa + p \log \mathbb{E}\left[\exp\left(\kappa Z_1^2 \mathbf{1}_{\{Z_1^2 > 1 + z_\star\}}\right)\right]\right]. \quad (18)$$

We calculate that

$$\mathbb{E}\left[\exp\left(\kappa Z_1^2 \mathbf{1}_{\{Z_1^2 > 1 + z_\star\}}\right)\right] = 2 \int_0^{\sqrt{1+z_\star}} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \mathrm{d}x + 2 \int_{\sqrt{1+z_\star}}^{\infty} \frac{e^{-\frac{1}{2}(1-2\kappa)x^2}}{\sqrt{2\pi}} \mathrm{d}x$$

$$\leq 1 + 2 \int_{\sqrt{1+z_\star}}^{\infty} \frac{e^{-\frac{1}{2}(1-2\kappa)x^2}}{\sqrt{2\pi}} \mathrm{d}x$$

$$\leq 1 + \frac{\exp\left(-\frac{z_\star(1-2\kappa)}{2}\right)}{1 - 2\kappa},$$

where the last inequality uses some easy algebra and the well known bound on the Gaussian cdf: $\int_t^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2a^2} \mathrm{d}x \leq \frac{a^2 e^{-t^2/2a^2}}{t\sqrt{2\pi}}$, valid for all $t > 0$. With $z_\star = c_0 \log p$, We deduce that

$$p \log \mathbb{E}\left[\exp\left(\kappa Z_1^2 \mathbf{1}_{\{Z_1^2 > 1 + z_\star\}}\right)\right] \leq p \log\left(1 + \frac{p^{-\frac{c_0(1-2\kappa)}{2}}}{(1-2\kappa)}\right) \leq \frac{p^{c_0\kappa}}{1 - 2\kappa}.$$

Hence with $a = \frac{2-2\kappa}{\kappa} \frac{p^{c_0\kappa}}{1-2\kappa} \leq 4\kappa^{-1} p^{c_0\kappa}$, (18) gives

$$\mathbb{P}\left[\sum_{j=1}^{p} Z_j^2 \mathbf{1}_{\{Z_j^2 > 1 + z_\star\}} > a\right] \leq \exp\left(-a\kappa + \frac{p^{c_0\kappa}}{1 - 2\kappa}\right) \leq \exp\left(-p^{c_0\kappa}\right).$$

We conclude that with probability at least $1 - \exp(-p^{c_0\kappa})$, $\sum_{j=1}^{p} Z_j^2 \mathbf{1}_{\{Z_j^2 > 1 + z_\star\}} \leq a \leq 4\kappa^{-1} p^{c_0\kappa}$, so that

$$\sum_{j \notin I_{\gamma_\star}} \left(\tilde{\beta}_{n,j} - \beta_{\star,j}\right)^2 \leq \frac{4\sigma^2}{cn} \kappa^{-1} p^{c_0\kappa}, \quad (19)$$

with probability at least $1 - \exp(-p^{c_0\kappa})$. Combining (17) and (19) it follows that

$$\|\tilde{\beta}_n - \beta_\star\|_2^2 \leq \frac{4\sigma^2}{cn}\left((1+c_0)s\log p + \frac{p^{c_0\kappa}}{\kappa}\right),$$

with probability at least $1 - \frac{1}{p^{c_0 s/8}} - \exp(-p^{c_0\kappa})$. Finally since $\log s > 1$, we can take $\kappa = \log(s)/(c_0 \log(p)) \in (0, 1/2)$ to achieve $s = p^{c_0\kappa}$. With this choice,

$$\frac{p^{c_0\kappa}}{\kappa} = \frac{s\log(p)}{c_0\log(s)} \leq \frac{s\log(p)}{c_0},$$

and the theorem follows easily.

$\square$

4.4. Proof of Corollary 5.

Proof. Recall that $\mathsf{I} = \mathsf{I}_{\gamma_\star} = \{1 \leq j \leq p : \beta_{\star,j} \neq 0\}$. We write $u_{\mathsf{I}_{\gamma_\star}} = (u_j, \ j \in \mathsf{I}_{\gamma_\star})$, and $u_{\mathsf{I}_{\gamma_\star}^c} = (u_j, \ j \notin \mathsf{I}_{\gamma_\star})$. It is clear that whenever (11) holds and $|\tilde{\beta}_{n,j} - \beta_{\star,j}| \leq \sqrt{M\sigma^2 s \log(p)/n}$, we have $\mathsf{sign}(\tilde{\beta}_{n,j}) = \mathsf{sign}(\beta_{\star,j})$. Since $|\tilde{\beta}_{n,j} - \beta_{\star,j}| \leq \|\tilde{\beta}_n - \beta_\star\|_2$, we conclude that $\mathsf{sign}(\tilde{\beta}_{n,\mathsf{I}_{\gamma_\star}}) = \mathsf{sign}(\beta_{\star,\mathsf{I}_{\gamma_\star}})$, with probability at least $1 - \frac{1}{p^{(c_0 s)/8}} - \frac{1}{\exp(s)}$.

For the other part, it follows from the definition of $\tilde{\beta}_n$ that for $\beta_{\star,j} = 0$, $\mathsf{sign}(\tilde{\beta}_{n,j}) \neq 0$ implies that $Z_j^2 \geq 1 + z_\star$. Hence

$$\mathbb{P}\left(\mathsf{sign}(\tilde{\beta}_{n,\mathsf{I}_{\gamma_\star}^c}) \neq \mathsf{sign}(\beta_{\star,\mathsf{I}_{\gamma_\star}^c})\right) \leq \sum_{j=1}^p \mathbb{P}\left(Z_j^2 > 1 + z_\star\right) = \sum_{j=1}^p 2\mathbb{P}\left(Z_j > \sqrt{1+z_\star}\right)$$

$$\leq \sum_{j=1}^p e^{-\frac{1}{2}(1+z_\star)} \leq \exp\left(\log p - \frac{c_0}{2}\log p\right)$$

$$\leq \frac{1}{p^{\frac{c_0}{2}-1}}.$$

The results follows.

$\square$

4.5. Proof of Proposition 6.

Proof. We fix an arbitrary $j \in \{1, \ldots, p\}$. We define $x_k = \|x_j\|^2 \gamma_j^{(k)}$, where we omit the dependence on j to keep the notation simple. It follows from (13) that

$$x_{k+1} = B\left(\frac{x_k}{\sigma^2 + x_k}\right)^2 + \frac{\sigma^2 x_k}{\sigma^2 + x_k} = \Psi(x_k),$$

where $B = \langle y, X_j \rangle^2 / \|x_j\|^2$, and

$$\Psi(x) = B\left(\frac{x}{\sigma^2 + x}\right)^2 + \frac{\sigma^2 x}{\sigma^2 + x}.$$

Notice that for all $x \geq 0$, $\Psi(x) \in [0, \sigma^2 + B]$. Hence the sequence $\{x_k, \, k \geq 0\}$ is bounded. The equation $\Psi(x) = x$ is equivalent to $x^2(x + \sigma^2) = x^2 B$. If $B \leq \sigma^2$, $\Psi(x) = x$ has a unique solution $x = 0$. If $B > \sigma^2$, then $\Psi(x) = x$ has two solutions $x = 0$ and $x = B - \sigma^2$. The derivatives of Ψ are given by

$$\Psi'(x) = \frac{\sigma^4}{(\sigma^2 + x)^2} + \frac{2x\sigma^2 B}{(\sigma^2 + x)^3}, \quad \Psi''(x) = \frac{-2\sigma^2 x(\sigma^2 + 2B) + 2\sigma^4(B - \sigma^2)}{(\sigma^2 + x)^4}.$$

We consider two cases

(1) Case 1: $B \leq \sigma^2$. Then $\Psi''(x) \leq 0$ for all $x \geq 0$. Hence Ψ is concave. This implies that for all $x \geq 0$,

$$\Psi(x) \leq \Psi(0) + \Psi'(0)x = x.$$

This implies that $x_k = \Psi(x_{k-1}) \leq x_{k-1}$. This means that the sequence $\{x_k, \, k \geq 0\}$ is bounded and non-increasing, hence has a limit $x_\star$. By continuity of Ψ, the limit point $x_\star$ satisfies $\Psi(x_\star) = x_\star$. Hence $x_\star = 0$, since we have seen above that 0 is the only fixed-point of Ψ when $B \leq \sigma^2$.

(2) Case 2: $B > \sigma^2$: Then $\Psi''(0) > 0$, and by Taylor expansion, in a neighborhood of 0, we have $\Psi(x) \geq \Psi(0) + \Psi'(0)x = x$ for all $x > 0$ small enough. If $x_\star = B_j - \sigma^2$ denotes the unique positive fixed point of Ψ, we can conclude that for all $x \in [0, x_\star]$, $\Psi(x) \geq x$, and for $x > x_\star$, $\Psi(x) < x$. Therefore, if $x_0 \in [0, x_\star]$, then $\{x_k, \, k \geq 0\}$ is increasing and bounded, hence converges to the unique positive fixed point $x_\star$ (recall that $x_0 > 0$). Whereas, if $x_0 > x_\star$, then $\{x_k, \, k \geq 0\}$ is decreasing and bounded, hence converges to the unique positive fixed point $x_\star$.

$\square$

REFERENCES

BALAKRISHNAN, S. and MADIGAN, D. (2010). Priors on the variance in sparse bayesian learning: the demi-bayesian lasso. In *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger* (M.-H. Chen, P. Muller, D. Sun and K. Ye, eds.). Springer, New York.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data.* Springer Series in Statistics, Springer, Heidelberg. Methods, theory and applications.

Faul, A. C. and Tipping, M. E. (2002). Analysis of sparse bayesian learning. *Advances in Neural Information Processing Systems* **14** 383–389.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22.
URL http://www.jstatsoft.org/v33/i01

Golub, G. and Van Loan, C. F. (2013). *Matrix Computations, 4th Ed.* John Hopkins University Press, Baltimore.

Meinhausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal.* **4** 85–117.

Teicher, H. (1984). Exponential bounds for large deviations of sums of unbounded random variables. *Sankhyā Ser. A* **46** 41–53.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58** 267–288.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *JMLR* **1** 211–244.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* Springer Series in Statistics, Springer-Verlag, New York. With applications to statistics.

Wipf, D. P. and Rao, B. D. (2004). Sparse Bayesian learning for basis selection. *IEEE Trans. Signal Process.* **52** 2153–2164.